Identification of potential RNA substrates for the 3'-5' polymerase BtTLP with RNA-Seq

SENIOR HONORS THESIS

Presented in partial fulfillment of the requirements for graduation with honors research distinction in Biomedical Engineering at The Ohio State University

By

Spencer Gardner

Undergraduate Program in Biomedical Engineering

The Ohio State University

2015

Committee:

Jane Jackman (Project Advisor)

Ralf Bundschuh (Project Advisor)

Jun Liu (Engineering Advisor)

Abstract

Reverse (3'-5') polymerases are a relatively new discovery and are found in all three domains of life. The *in vivo* function of many of these proteins remains ambiguous. BtTLP, a reverse polymerase from the soil bacterium, *Bacillus thuringiensis*, has recently come under investigation through structural, genetic, and biochemical analysis. The overall goal of this study is to develop a new form of RNA-Seq to identify potential substrates for BtTLP in an engineered system (a previously characterized strain of *Saccharomyces cerevisiae* (baker's yeast) expressing BtTLP), with the ultimate goal of elucidating a more complete understanding of these unusual 3'-5' polymerases in biology. This project is being accomplished both from the wet and computational aspects. To start, a library of all small RNAs (approximately less than 200 bp) from the engineered system has been generated and sequenced. A computational pipeline has been developed to process the large amount of data that comes from deep-sequencing experiments. It is expected that multiple new substrates will be identified based on previous *in vitro* biochemical studies and an observed growth defect in the engineered system when compared to an isogenic control. This approach has significant advantages over the laborious alternative of testing each RNA in the cell individually. Traditionally, RNA-Seq has been used to identify functional elements in the genome, but here it is being used to detect post-transcriptional 5' nucleotide addition.

Introduction

Essentially all canonical DNA or RNA polymerases synthesize nucleic acids in the 5' to 3' direction where the 3' hydroxyl group on the growing nucleic acid performs a nucleophilic attack on the α-phosphate group on the incoming nucleotide triphosphate (NTP), liberating the β and γ phosphates as a pyrophosphate molecule[7]. This mechanism could, in principle, work in the same manner but with the relevant players switched, where the 3'-hydroxyl of the incoming NTP performs a nucleophilic attack on the growing 5'-triphosphorylated end of the nucleic acid, thus effecting RNA/DNA synthesis in the opposite (3'-5') direction[7] (figure 1). Yet, in all biological systems discovered to date, DNA and RNA polymerases appeared to universally catalyze the 5'-3' reaction, and this was thought to be the only mode of synthesis used in biology.

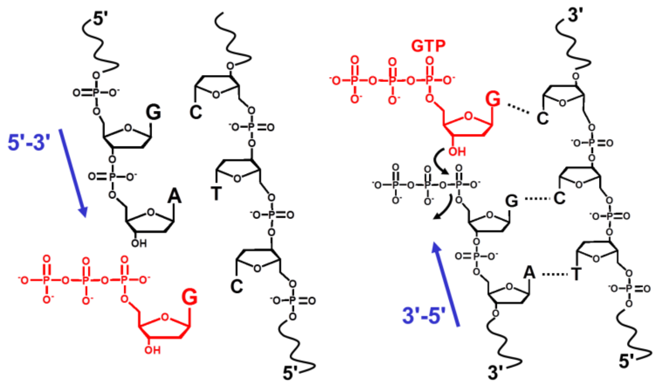


Figure 1: [Left] Shows the action of all previously known DNA/RNA polymerases. [Right] Shows the action of tRNAHis guanylyltransferase (Thg1) Thg1-like proteins (TLPs).

This picture changed when a family of reverse polymerases was discovered[6,7]. The initial reverse polymerase was discovered in *S. cerevisiae* and adds a nontemplated G_{-1} across the A_{73} discriminator nucleotide in $tRNA^{His}$, and was named $tRNA^{His}$ guanylyltransferase (Thg1)[6,7]. Initially, this *in vivo* action of Thg1 was not obviously connected with that of a reverse polymerase since it only involved the addition of a single nucleotide[6,7]. However, when the discriminator A_{73} nucleotide associated with the biological $tRNA^{His}$ substrate was changed to a C_{73}, Thg1 catalyzed the addition of multiple guanine residues using the 3' portion of the acceptor stem as a template[6,7]. Since the initial discovery of this reverse polymerase activity, proteins that are homologous to *S. cerevisiae* Thg1 have been identified from all domains of life and are classified into two subfamilies of enzymes called Thg1 and Thg1-like-proteins (TLPs)[1,7]. TLPs prefer to add nucleotides that form traditional Watson-Crick base pairs[1,2,9,10]; thus Thg1 is the only reverse polymerase in this family that also efficiently adds a non-Watson-Crick base pair[6,7] (figure 2).

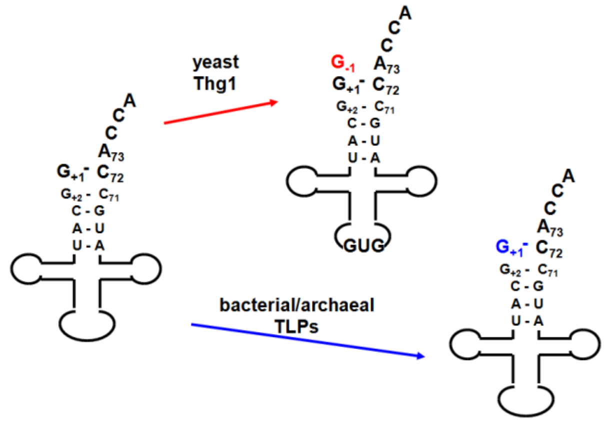


Figure 2: Shows the preferred action of Thg1 and TLPs where Thg1 prefers to add a non-templated guanine residue while TLPs prefer to form standard Watson-Crick base pairs.

In vivo TLP characterization has proven difficult due to a number of technical challenges, and thus only one bona fide biological activity has been associated with TLPs[2,7,9]. However, extensive *in vitro* characterization of multiple purified TLPs (from Bacteria, Archaea and Eukarya), has shown these enzymes to act on various 5'-truncated tRNAs, so many of these TLPs may play a role in various tRNA quality control pathways[2,7]. This *in vitro* observation was applied to investigation of the TLP enzymes that are encoded in the eukaryotic slime mold *Dictyostelium discoideum*, where one enzyme (known as DdiTLP3) has been demonstrated through both biochemical and genetic experiments to utilize this activity to participate in a process known as mitochondrial tRNA 5'-editing[2]. Importantly, this result not only substantiated the first physiological role for 3'-5' polymerase activity, but also opened the door to identification of additional roles for these enzymes, including in *D. discoideum*.

The TLP from *Bacillus thuringiensis*, known as BtTLP, which is the focus of this work, has also been studied previously[7,9]. Biochemically, *in vitro* data has shown BtTLP to act on a variety of RNA substrates from simple hairpins to large RNAs, but the host *in vivo* function is still unknown for BtTLP, or for that matter, any archaeal or bacterial TLP[7]. It is known that BtTLP does not share the same *in vivo* function as the TLPs from *D. discoideum* because the mitochondrial tRNA 5'-editing reaction does not take place in *B. thuringiensis*[7,9]. In addition, *B. thuringiensis* contains a genomically encoded G_{-1}, raising questions about whether or not RNase P removes that nucleotide along with the rest of the precursor sequence and if so, what BtTLP is doing in its native organism. To study BtTLP, it would be desirable to develop a genetic system in *B. thuringiensis* but there are no genetic tools available to investigate the function of BtTLP in its native bacteria, i.e., a genetic knockout has not been possible[9]. Thus to address the issue in another way, genetic systems were developed in yeast to try to detect activities of this enzyme expressed heterologously[9]. Preliminary analysis of this system revealed that BtTLP is functional in yeast, and that it may act on unique RNAs in a yeast cell[9] (figure 3).

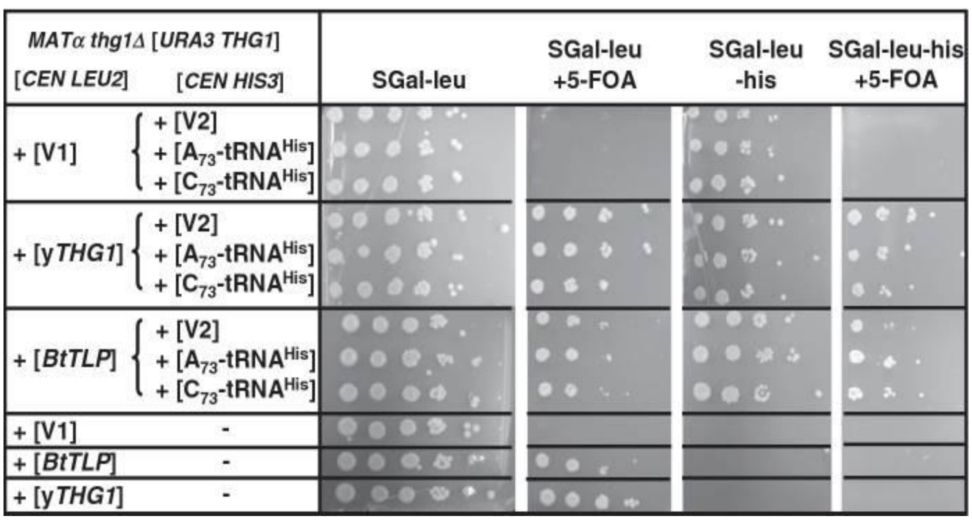


Figure 3: BtTLP has been shown to weakly compliment yeast Thg1 function. This could be due to BtTLP acting on other RNAs in yeast, other than tRNAHis, in a deleterious way.

The main goal of my project is to identify potential substrates in yeast for this enzyme through a variation on a relatively new RNA deep-sequencing technique known as RNA-Seq[11].

This will allow every small RNA (initially less than ~200 bp) in *S. cerevisiae* cells to be observed to see if it is a substrate for BtTLP.

Methodology and Results

Methodology overview

On the Biochemical end, the goal was to prepare a library of every small RNA in the cell (approximately less than 200 bp) from two strains of *S. cerevisiae*. The first strain was a wild type strain (with a wild type copy of the *THG1* gene on an extrachromosomal plasmid) that acted as an isogenic control. The second strain was a *BtTLP*-complemented cell line and acted as the experimental group. The goal was to compare differences in total small RNAs from the two cell lines.

The computational side required the assembly of what is known as a computational pipeline. The purpose of this pipeline was to process the massive amount of data that came from deep sequencing. This has been done by assembling programs that have been inherited from others in the Physics department; utilizing software packages such as STAR[3], Bowtie[8], and Samtools[14]; as well as writing multiples programs of my own to fill in the gaps. In addition, the UCSC Genome Browser[12] has been used to visualize data.

Plasmid shuffle assay

Three strains of yeast cells were used in this study. The starting strain was a parental strain used to derive the other two strains. The first strain (JJY240) acted as a negative control in this study. It contains chromosomal deletions of the *thg1*, *ura3*, and *leu2* genes. Without the *URA3* or *LEU2* genes the cell is unable to synthesize uracil and leucine, respectively. Thus if the cells are grown on media that does not contain uracil or leucine then the cells will not survive. In addition, *THG1* is an essential gene. The two strains derived from JJY240 were the positive control (JJY242) and the experimental strain (JJY247). All three strains contained a covering plasmid with the *URA3* and *THG1* genes, allowing the cells to survive on uracil drop out media. All three cell lines also contained a second plasmid with two genes on it. The first gene, *LEU2*, allowed the cells to survive on leucine drop out media. The second gene on this second plasmid was different in each of the three strains. In JJY240 this spot contained a galactose inducable promoter but with no gene, the promoter was empty. In JJY242 this spot contained the *THG1* gene with a galactose inducible promoter. In JJY247 this spot contained the *BtTLP* gene with a galactose inducible promoter (figure 4).

The three strains were streaked out from -80^{0}C onto rich media plates to get initial cell growth. The media consisted of yeast extract, peptone, and dextrose (YPD) at 37°C. Under those growth conditions, cells do not need to synthesize their own nutrients, such as uracil and leucine. Next, single colonies from each strain were plucked from the plates and streaked onto drop out media plates. This media contained synthetic dextrose and was missing uracil and leucine (SD-ura-leu). This made it so that any cells that had not retained both plasmids would not be able to survive. Then after this first round of selection, colonies from the SD-ura-leu plates were streaked out onto plates containing galactose, 5-Fluoroorotic Acid (5FOA), and were still missing leucine but did contain uracil (SGal-leu+5FOA). When 5FOA is present, it acts as a toxin to cells containing the URA3 gene[4]. Only cells that lost the plasmid containing the *URA3* and *THG1* genes are potentially viable. The galactose in the media induces the promoter on the second plasmid that had previously been inactive when dextrose was present. Under those conditions the JJY240 (negative control) cells all died because there was nothing to complement the *THG1* function that was on the lost *URA3* plasmid. The JJY242 cells (positive control) remained viable as there was *THG1* gene on the *LEU2* plasmid to complement the lost *THG1* on the *URA3* plasmid. JJY247 (experimental group) cells were viable as well, meaning the *BtTLP* gene was able to viably complement the lost *THG1* gene (figure 4).

Individual colonies were then removed from the plates and grown in 5ml yeast-peptone-galactose (YPGal) liquid media overnight at 37° C and then 100µl of culture was transferred to 500ml YPGal liquid media and grown for about 24 hours until the optical density (OD) at 600nm (OD_{600}) reached ~1. This was done to ensure that the cells were still in their exponential phase of growth.

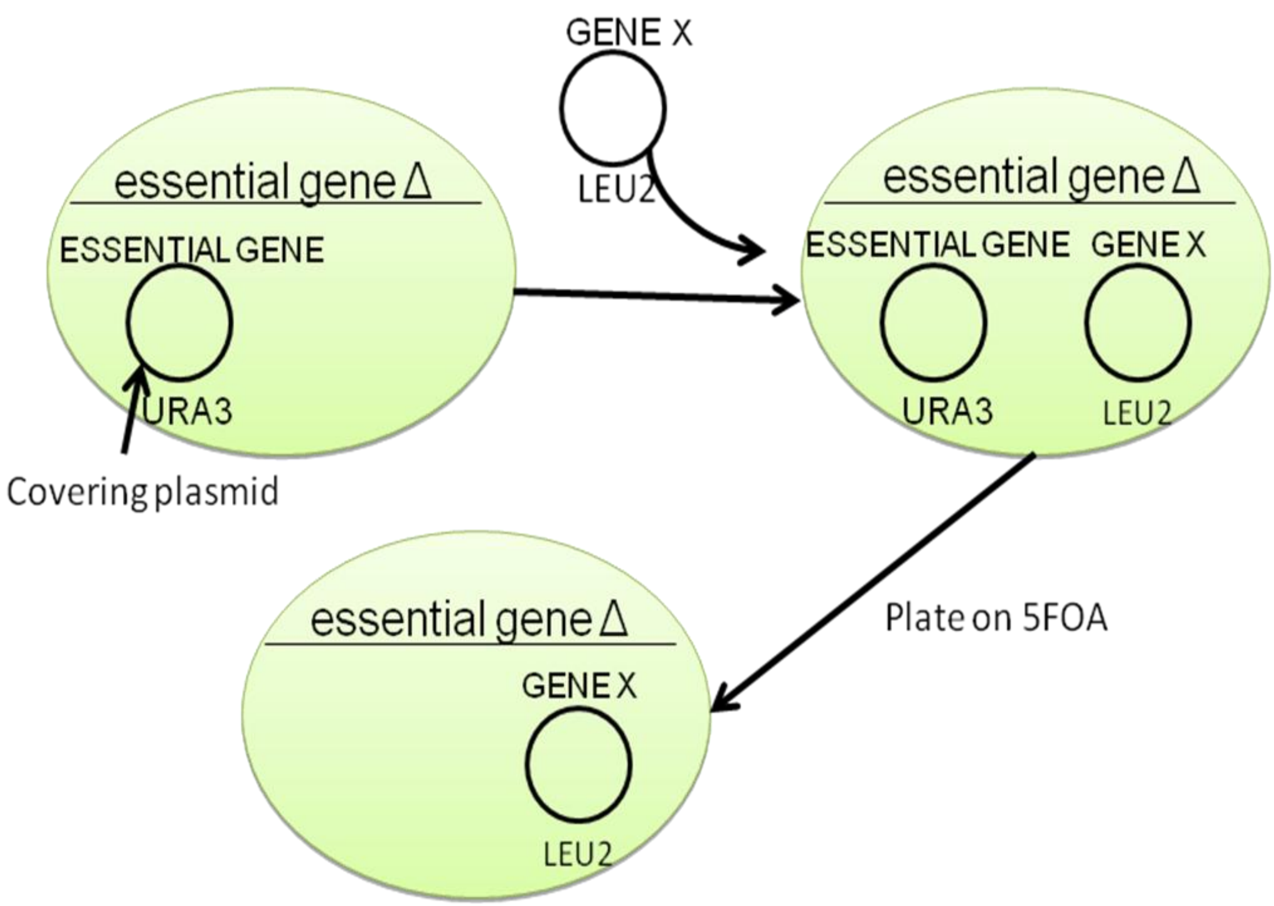


Figure 4: The general strategy of a plasmid shuffle assay. Note that leuΔ and uraΔ are not shown for simplicity but are present. Gene X varies between the three strains. JJY240 has nothing there, JJY242 has THG1, and JJY247 has BtTLP.

Cells were then harvested by centrifugation, washed with ice-cold sterile water, quickly frozen in dry ice for ~5min, and then stored at -80°C.

Cell lysis

Yeast pellets were removed from the -80°C freezer and allowed to thaw on ice. 300ODs of cells were then resuspended in RNA extraction buffer, consisting of 0.1M Sodium acetate (NaOAc), 20mM Ethylenediaminetetraacetic acid (EDTA), and 1% Sodium dodecyl sulfate (SDS). 300ODs refers to amount of cells in 300mL of culture that measures an OD_{600} of 1. To determine the best lysing method to preserve RNAs, several different methods were attemped. Yeast were lysed by a cell cracker (Isobiotech), by addition of phenol and vortexing, by addition of phenol and glass beads and vortexing, and with a French press. The lysed samples were then run through a PCA isolation and ethanol precipitation protocol that was aimed at retaining small RNAs. The recovered RNA was then resuspended in 0.5mL ddH_2O. A DNase treatment was also performed on the samples to verify that there was no DNA contamination. The extracted RNA derived from these four methods was resolved on an agarose- formaldehyde denaturing gel to qualitatively assess the purity, quantity and quality of the RNAs isolated by each method (example data shown in figures 5 and 6). The samples obtained by lysis with the cell cracker followed by phenol extraction had consistently degraded RNA, as evidenced by faint bands overall and one large band at the bottom of the gel. The samples obtained by lysis with the French press looked better qualitatively but still suffered from similar RNA degradation problems, meaning larger bands were more distinct but overall still suffered from one bright band at the bottom of the gel. Lastly, lysis by phenol and vortexing, and lysis by phenol and beads and vortexing had similarly positive results, consistently showing the least amount of

RNA degradation, with distinct bands that most likely correspond to the different rRNAs and a fainter band at the bottom of the gel than the other lysis techniques.

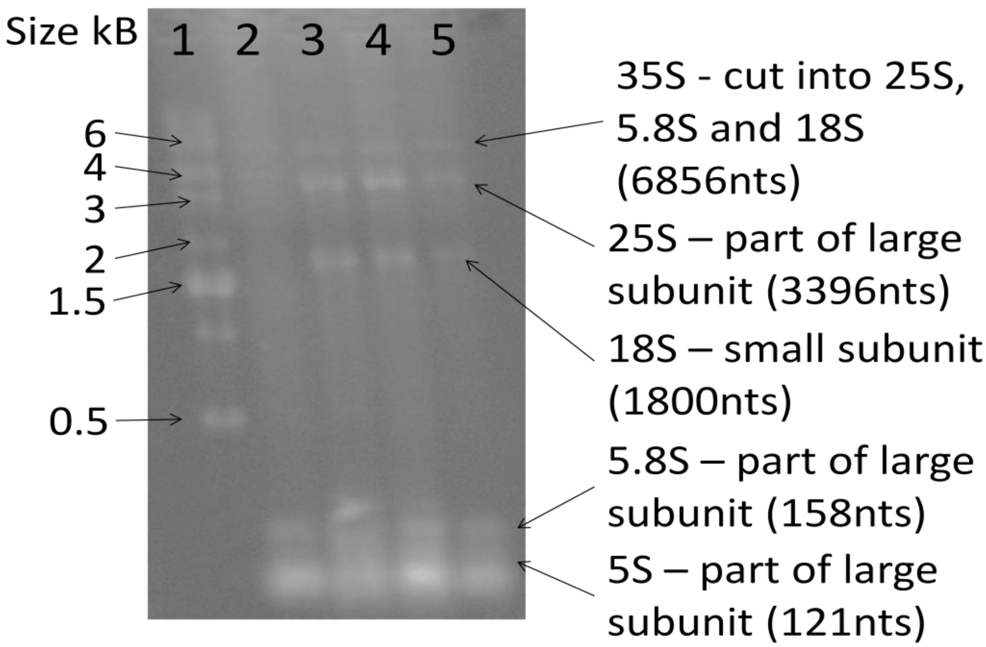


Figure 5: A denaturing agarose gel demonstrating total RNA samples with ribsosomal peaks to analyze the quality of isolated RNA. Lane 1 is a 0.5-10 kilobase (kB) ladder. Lanes 2 and 3 are the isogenic control. Lanes 4 and 5 are the experimental group.

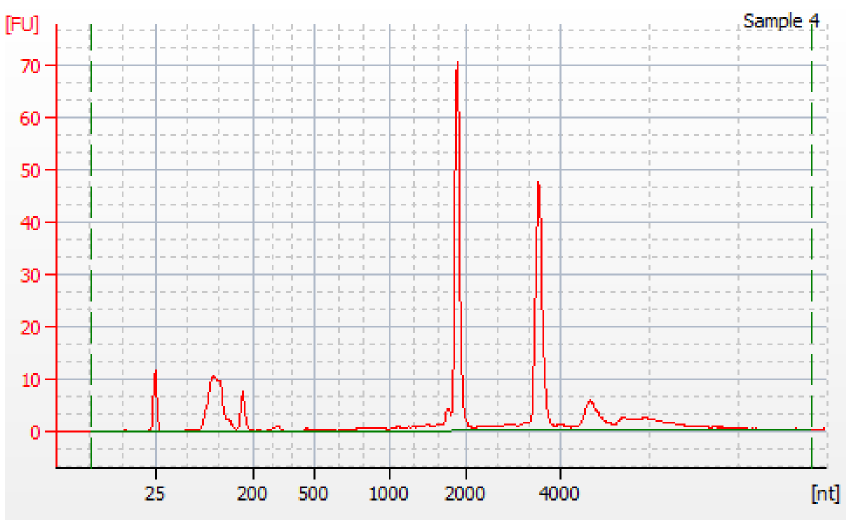


Figure 6: Bioanalyzer output shows distinct peaks, demonstrating a high quality RNA sample.

Small RNA isolation

In order to remove much of the larger ribosomal RNA (because the high abundance of these RNAs would cause them to comprise the major fraction of reads obtained in the RNA-seq experiment), the high quality samples from the previous step were run through the mirVana^{TM} miRNA Isolation Kit from Ambion® (figure 7). This kit is effective at separating the RNA into two fractions, one containing all RNAs approximately below 200nts (figure 7). The kit works by running the sample through two glass fiber filters. The sample is first passed through at a low ethanol concentration in order to remove larger RNAs. The sample is then passed through another filter at a high ethanol concentration to precipitate the remaining RNAs on the glass fiber filter. The RNAs are then eluted from the filter with ddH_2O and quantified using a Nanodrop and Infrared spectroscopy (IR).

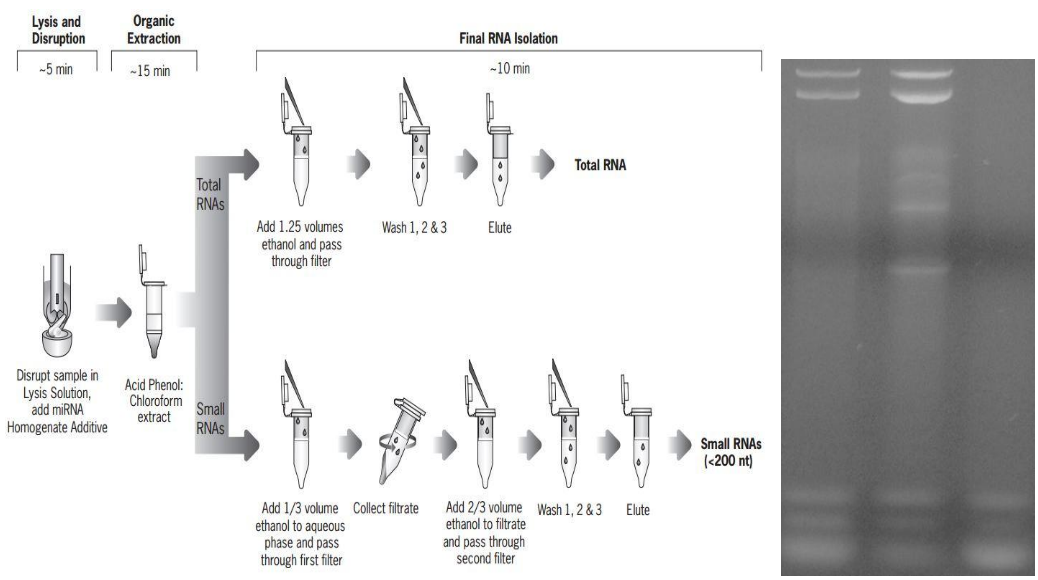


Figure 7: An overview of the size fractionation protocol. The RNA sample is first run through the glass fiber filter at a low EtOH concentration to remove large RNAs. The process is then repeated with a high EtOH concentration to remove the remaining RNA. The remaining small RNAs are then eluted from the glass fiber column. The image on the right demonstrates that the kit effectively separates an RNA sample into large and small fractions. Lane 1 is total RNA before fractionation. Lane 2 is the large fraction, note that while it was effective at removing large RNAs it also removed smaller ones as well. Lane 3 is the small RNA fraction, and it can be seen that there is virtually no large (ribosomal) RNA remaining.

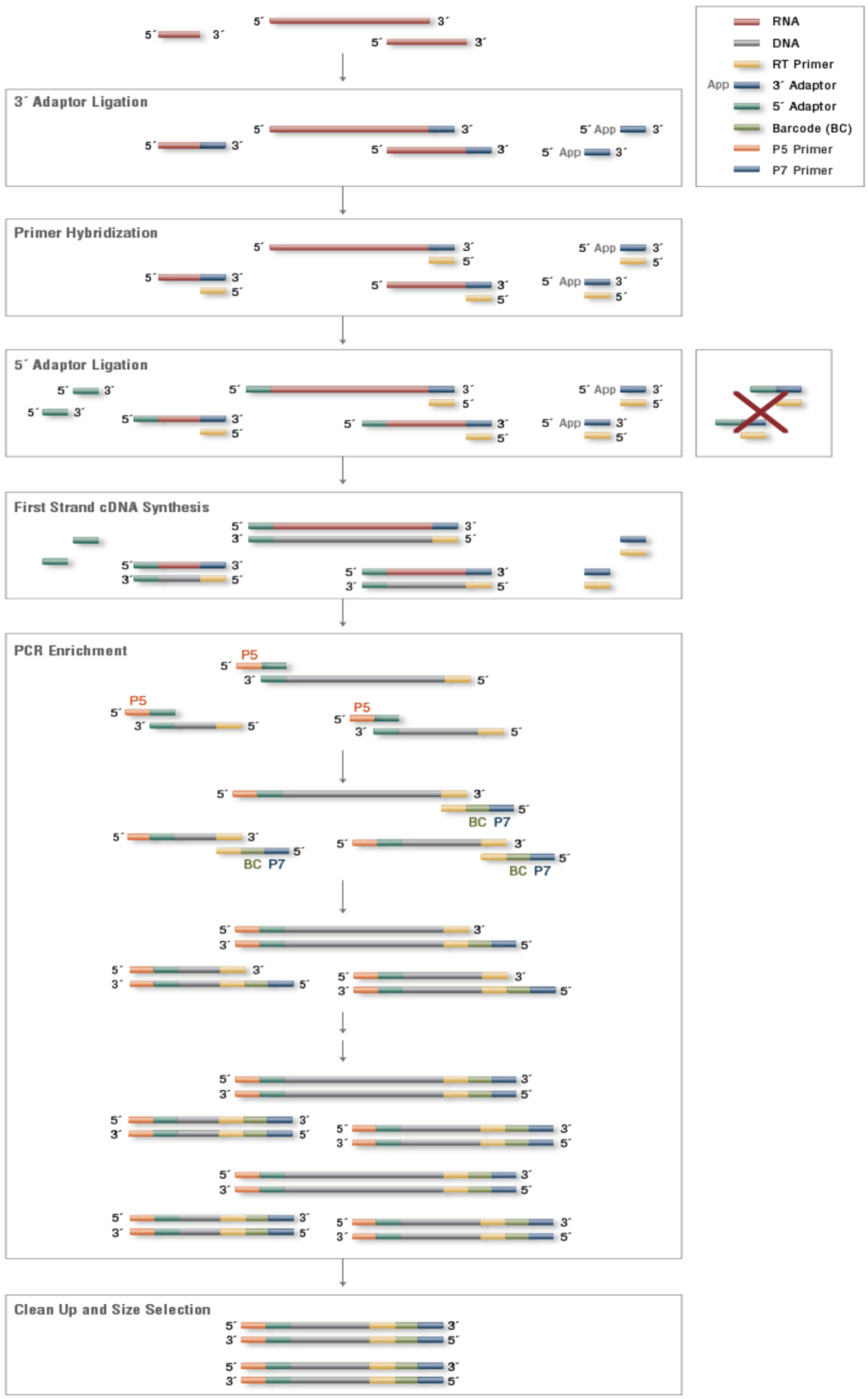


Figure 8: A summary of the steps in the library generation process. This Version of RNA-Seq differs from preparations in the past in that the adaptors bind directly to the 5' and 3' ends of the molecule as opposed to hybridizing to some identity element in the RNA sample, such as a poly-A tail with mRNAs.

Library preparation

In order for the adapters to be ligated properly in the library preparation, the RNAs must only contain only a monophosphate on the 5'-end. Incubation of the purified RNA with Tobacco Acid Pyrophosphatase (TAP) was used to remove any 5'-5' cap structures from the isolated RNA as well as any 5'-triphosphorylated ends that could result from a number of sources and yield 5'-monophosphorylated RNAs to be processed further.

The library was prepared with NEBNext® Small RNA Library Prep Set for Illumina (New England BioLabs inc.) according to the manufacturer's instructions. Briefly, the steps include 3' adapter ligation, primer hybridization, 5' adapter ligation, 1st strand cDNA synthesis, and PCR amplification of cDNA (figure 8).

The amplified cDNA was then analyzed by RNA-Seq at The Ohio State University Cancer Center. Raw data obtained from sequencing was treated according to the computational pipeline described below.

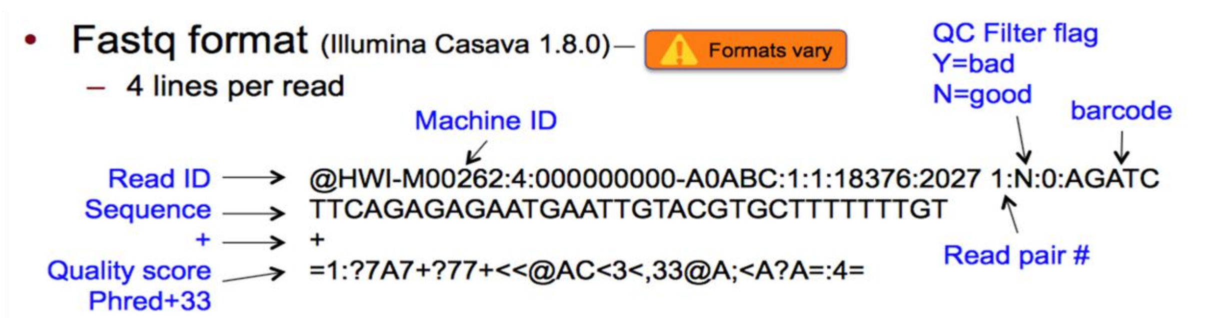
Computational pipeline overview

In order to filter through the millions of reads that come from deep sequencing data a computational pipeline was developed. Figure 9 summarizes the series of programs and software that the data is run through. The entire pipeline is run on a Unix system and some programs require the use of MATLAB to run while others are written in C or awk.

Initial data generation and primary analysis

To start with, for the sake of validation, a program that generates artificial sequence data was used initially. This allows the user to build in certain trends to verify the pipeline at the end.

The program works by feeding in cDNA sequences and then taking a user determined number of reads, read length, and error rate and generates a FASTQ output file. This program will also add the adapters that are used in RNA-Seq in order to simulate real data as much as possible.



1. Generate sequence information (Illumina RNA sequencing)
2. Remove low quality reads and adaptors from FASTQ data files
3. Trim nucleotides from 5’ end of read from FASTQ data files
4. Align reads to reference genome (SacCer_Apr2011/sacCer3)
5. Separate reads in forward and reverse strands
6. Add 10 positions back to each read and convert to BED format to properly visualize downstream
7. Identify 5’ ends (peaks) in isogenic control data
8. Sum reads in 5’ direction to locate post transcriptional nucleotide addition
9. Calculate statistical significance between summed values between control and experimental values to make a list of potential targets to look at experimentally

Figure 9: [Top] demonstrates the general FASTQ file format. This is the format that data is sent out from the sequencer. It is the raw sequence information. It contains several important features, a header, the sequence itself, and then a quality score associated with each base. There were approximately ten million of these reads for each of the samples run in this study. [Bottom] is the general data analysis pathway run through various computational steps.

For the actual experiment, Illumina sequencing of the RNA-Seq library was conducted. This is a complicated process and some of the details associated with this type of next-generation sequencing are proprietary. Nonetheless, the major steps associated with this type of sequencing technology are described. Samples are washed over a flow cell, which is a glass slide with lanes. Each lane is covered in two different oligonucleotides (oligos). The first oligo is able to hybridize with the adaptor region of all reverse strands in the sample. A polymerase then creates a complement of the hybridized fragment. This double stranded DNA molecule is then denatured and the original strand is washed off. The remaining strand is then hybridized with the adaptor region on the other end of the molecule to the second oligo, creating a bridge that a polymerase uses to create a complementary double stranded molecule. The molecule is then denatured again, resulting in two molecules, each one bound to a different oligo, creating forward and reverse strands. This bridge amplification process is repeated many times. This happens simultaneously for every DNA molecule from the sample, creating millions of these clusters on the flow cell. Once this bridge amplification process is completed, the reverse strands are cleaved and washed away so that only the forward strands remain bound to the flow cell. The 3' ends are then chemically blocked off to prevent priming during sequencing. The 5' end of the molecules are then hybridized with a sequencing primers and fluorescently tagged nucleotides are sequentially added to the DNA molecule. After each nucleotide is added, light excites the newly added NTP and it gives off a distinct color depending on the identity of the added base. This process is called sequencing by synthesis and the specifics of the process are unfortunately proprietary. This fluorescent process takes place with every molecule on the cell at once in a massive parallel process. After each read, the new strand is washed away and the process is repeated. For paired-end sequencing, as was obtaining in this study, a complementary DNA molecule is synthesized and its 3' end blocked. Sequencing then proceeds in the same manner as the first read, but this time sequence information is obtained from the other end of the molecule at the start of the read. This is useful in determining the identity of ambiguous reads. The current pipeline is not developed to handle these paired-end reads, thus only half of the information is currently useable. Future work will aim to adapt the computational pipeline to incorporate paired-end reads.

The output (FASTQ) files are then fed into a program that removes low quality reads based on some thresholds defined by the user. In addition, this program trims the adapter sequences that were introduced in the previous step.

Secondary analysis

High quality reads are then fed into a program written in MATLAB that cuts off the last 10 bases from the 5' end of each read. This must be done in preparation for the alignment step. The goal of this project is to identify post-transcriptional nucleotide addition on the 5' end of RNA molecules, but oddly enough the presence of these additional 5' nucleotides will prevent the reads from properly aligning to the reference genome. In addition, it is unknown which potential RNA molecules will be substrates for the enzyme of interest, nor is it known how many nucleotides will be added. Experimental data so far shows only a few nucleotides to be added on known RNA substrates. Based on that, a conservative number of 10 nucleotides are removed from each read to aid in the alignment process.

Chopped reads are then fed into a software that takes the reads and aligns them to a reference genome. The reference genome for yeast was recently updated by the *Saccharomyces Genome Database* project in April 2011, and that is the reference that was used in this study. The Ultrafast Universal RNA-Seq Aligner (STAR) was the software used to align reads in this study[3]. STAR has the benefit of accurately aligning spliced reads *de novo,* which is the case in some tRNA genes in Yeast. This is an issue that other alignment software's are unable to accomplish, such as Bowtie[8].

From there, Samtools separated the data into forward and reverse strands. Samtools is a software that manipulates sequencing information. STAR outputs data into the standard SAM (Sequence Alignment/Map) format, which is generic file type for storing large nucleotide sequence alignment information. Samtools then can manipulate SAM files to sort, index, and organize them in various ways. Once the sequences are separated into forward and reverse strand data files, a program turns the data into BED files (figure 10). BED files are a simple numerical way to organize the data. Each read has information about which chromosome the read aligned to and where the read aligned to on that chromosome. In addition, there is a read count for the number of identical reads at a given position. In the following steps these BED files are run through different programs written in MATLAB to search for potential 5' nucleotide addition.

chrII	326865	326865	1.44E+03
chrII	643078	643078	1.42E+02
chrIII	123648	123648	7.18E+02
chrIV	802802	802802	2.15E+02
chrIV	1075544	1075544	7.24E+02

Figure 10: This is typical BED file format, albeit truncated for the sake of visualization. Each row corresponds to a unique read. The first column is the name of the chromosome that the read aligned to. The second and third locations are the start and stop points on the given chromosome. Although in the study only the 5' end of each read is under investigation so the start and stop points are same because each read only consists of one nucleotide. The forth column is the number of reads that mapped to that location. These values are what are summed when looking for peaks beyond the 5' end.

Figure 11 shows BED files that have been uploaded to the UCSC Genome Brower. Since BED files only contain information on the 5' end, each read shows up as a peak rather than a bar. Higher peaks means there are more reads at that position. Thus a goal of this pipeline is to identify peaks that may correspond to BtTLP nucleotide addition by searching for peaks in the BtTLP cell line data that are not present in the Thg1 cell line data. It should be noted that the following steps are run for forward and reverse strands on the experimental (BtTLP) data and the control (Thg1) data. This means that for one replicate of experimental and control data these steps are run four times. Overall what has been done is the control data was used to create a list of peaks, then the experimental data was searched beyond those peaks to see if there has been any shift that could correspond to 5' nucleotide addition.

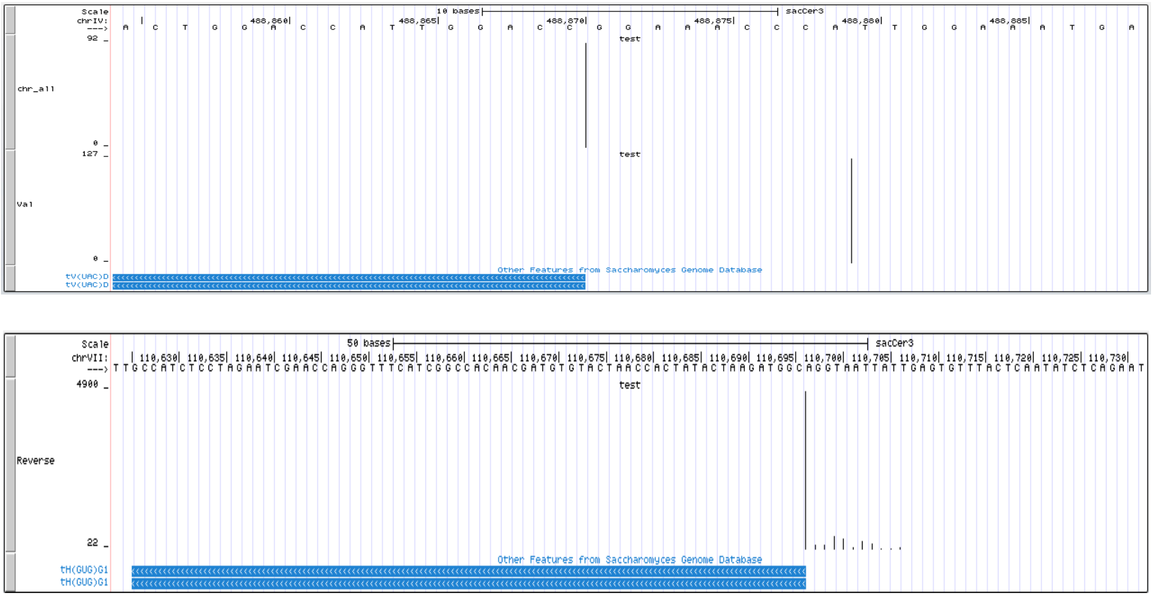


Figure 11: The two screen shots above show two BED files uploaded to the UCSC genome browser. The cases shown here are not real data, but were generated to demonstrate what the ideal nucleotide addition case would look like. [Top] shows two peaks. The first peak on the left shows the 5' end of reads that correspond with the 5' end of a tRNA^{Phe} gene. The peak to the right shows a case when exactly ten nucleotides were added and thus there was a shift in the 5' location of those reads. [Bottom] shows the 5' end of the same tRNA^{Phe} but with a smear of peaks. This more accurately resembles what real sequencing data looks like. Most of the reads are at the 5' end of the gene with a smaller number peaks reaching beyond the 5' end.

Tertiary analysis

Each BED file of control replicate data was added together so that peaks that may not be in every replicate were still present. Also, adding the BED files together helped to better define peaks and minimize the presence of sequencing artifacts. Then with files added together, all reads below some user defined threshold were removed. This was done to try and remove artifacts that were downsized in the first step. From there, each peak went through a series of if statements asking about the size of the peak relative to neighboring peaks to determine if the peak corresponded to the 5' position of a RNA molecule. Lastly, all peaks that met the criteria for 5'-end peaks were added to a list so that those positions could be searched in the experimental and isogenic control data and summed around those positions to search for statistical difference between the two groups. Figure 12 shows a flow chart for the criteria of a peak.

The files to be searched were first normalized. This was done by taking each read number and dividing it by the total number of reads. Each read number was then multiplied by some standard number to make the values easier to work with. Normalizing allows data from different replicates to be compared.

Once each replicate had been normalized they were searched according to the potential peaks list generated previously. This was done by going through each position in the potential peaks list and then finding that peak in the experimental data and then summing together the next few positions on the experimental and control data. If the values between the two data sets were statistically different then it would be said that 5' nucleotide addition has taken place. Further experimental study will then be warranted to validate the finding.

Alignment summary

Each sample aligned with similar results. There were approximately 10 million reads. Only 1.47% of reads were uniquely mapped. Once the adaptors were trimmed the average read length was only 38 bases. 93% of reads mapped to ten or less loci and less than 1% of reads were thrown out for matching more than 10 times. The mismatch rate and indels were also well under 1%. Overall the mapping statistics demonstrated high quality information.

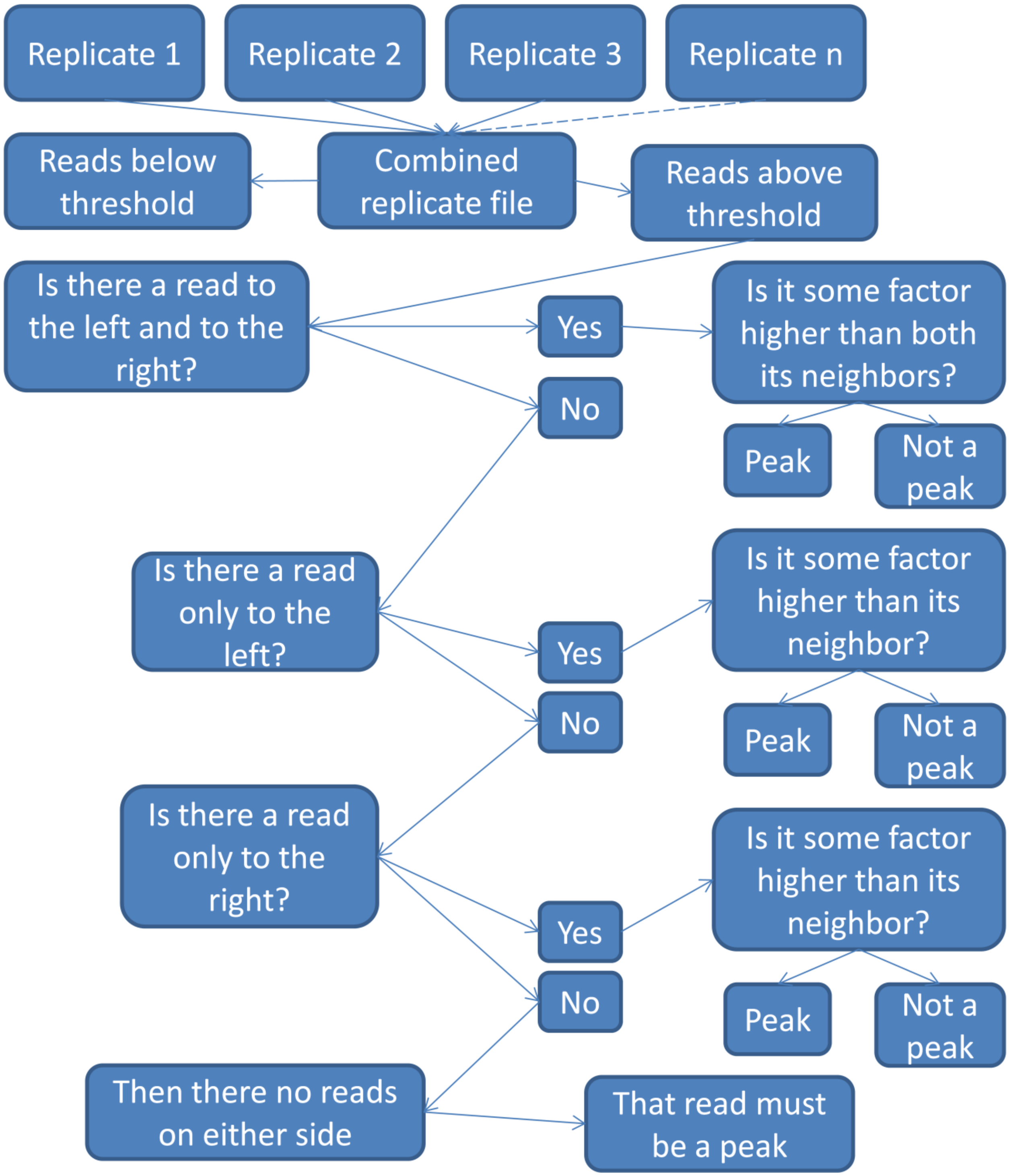


Figure 12: A flow chart showing how peaks are located. Each row of the BED file goes through the flow chart. There are four potential cases that a read can meet and be a peak (summarized above). Rows corresponding to peaks are pulled out and assembled into their own peaks BED file.

tRNA examination

Based on the well-known function of Thg1, it was of interest to directly look at different tRNA sequences to see if an A_{-1} was added to $tRNA^{His}$ in the control group and then to see what was added in the experimental group. Comparing the primary sequences obtained from $tRNA^{His}$, a difference in the -1 position was seen. As expected, the *Thg1*-complemented sequencing information showed that a G_{-1} nucleotide was added post-transcriptionally to this tRNA. However, the experimental strain expressing BtTLP contained a U_{-1} nucleotide on nearly all the $tRNA^{His}$ sequences. This finding makes sense in light of previous data because it is already known that TLPs prefer to make Watson-Crick base-pairs while Thg1 is the only enzyme that prefers to add a G_{-1} opposite of an A_{73}[1,6,7]. This may contribute to the observed growth defect as it has been shown *in vitro* that the HisRS complex it slowed 10 fold when G_{-1} is changed to U_{-1}[9]. Although given the other justifications for this project, it is plausible that BtTLP is acting on other RNAs to contribute to the observed growth defect.

$tRNA^{Gln}$ also demonstrated a small number of Watson-Crick single nucleotide addition as well. However, the number of sequences that contained additional nucleotides was small compared to the total number of sequences obtained for this RNA.

Discussion

Limitations

The way the computational pipeline is set up, visualization of peaks on the UCSC Genome Browser does not allow the user to differentiate between peaks in the same location that vary in sequence. This is because the BED file format only has positions, not sequences. Thus, once the sequence information is converted from a .bam file to a BED file the sequence identity is lost. BED files are much easier to work with though in identifying differences in peaks between samples.

All reads do not align as expected (figure 13). Parameters that STAR runs on will need to be optimized. Also, corroborating evidence about the true alignment could come from comparing information between different aligners, which may prove insightful for reads mapped where no gene is known to be.

The current libraries each contained approximately 10 million reads. Increasing sequencing depth will aid in teasing out differences between relatively low abundance RNAs. A previous study showed BtTLP to add nucleotides to different tRNAs and related structures *in vitro* and those products were not all identified in the current sequence information.

RNA-Seq library generation also introduces unknown biases into a pool of total RNAs. Hopefully over time these biases will be identified and accounted for in the literature.

Parameters

Various parameters are used throughout the computational pipeline. It is not known what the best set of parameters are and optimizing them in future studies may yield additional results. First, the program that removes low quality reads and trims adaptors has a set of parameters to determine what a low quality read is and when it should be removed. Each base must have a quality score of 20, meaning there is a 0.01% maximum uncertainty of the identity of a base. The next parameter is the minimum number of bases from the adaptor sequence that must be seen before it is trimmed. 5 is the value currently being used. If too low a value is picked then statistically more sequences will be trimmed because they will randomly appear in the reads themselves. Although, this number should be kept low because short sequences will be read

Figure 13 part 1: Alignment works well with the vast majority of sequences aligning to the genome. [A] these reads did not align properly as there is no transcribed region at this location. A BLAST search of this sequence confirmed that it was not contaminating RNA. [B] tRNAs often do not sequence well as is shown here with $tRNA^{Gln}$. Often it can be seen that the reverse transcriptase was halted at a particular base. In this case it can be seen exactly where cDNA synthesizing machinery was held up.

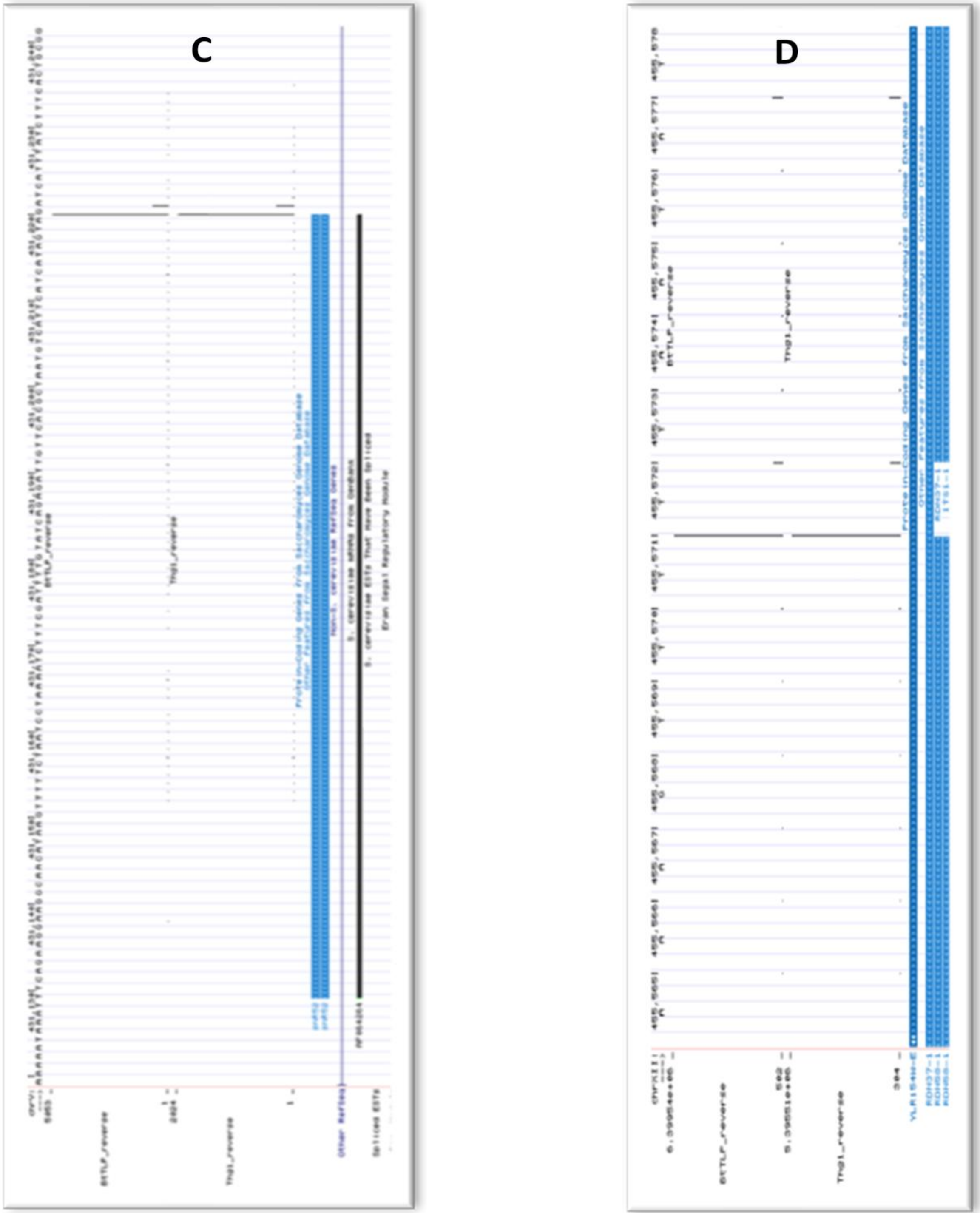


Figure 13 part 2: [C] in many cases interesting phenomena are noted such as here where an extra base appears to be at the end of this snRNA gene. Although the data between the control and experimental groups look identical. Increased sequencing depth and replicates will be needed to tease apart small differences in sequencing output. [D] shows the 5.8S rRNA gene. Note that half of the sequences align here. This may be acceptable though if it is not trivial to knock down the amount of small rRNAs in the sample. The reason for size fractionation was to remove the larger rRNAs, which would have otherwise swamped all other reads in the sample.

through entirely and then on past the sequence into the adapter on the opposite end. Not properly removing adaptor sequences throws off the next step of the pipeline where the remaining high quality reads are aligned back to the yeast genome. In addition, to allow for sequencing error, one base of the adaptor sequence is allowed to be errant, meaning only 4 out of the 5 bases must be correct. Next, the minimum sequence length after trimming the adaptors is 20 nucleotides. This value may need to be lowered in the future because it was noted that some sequences were consistently showing up that were smaller than 20 nts. The last parameter for this program determines the number of uncertain bases allowed in a single read. Currently this number is set at 5. Decreasing this number may increase alignment fidelity.

Once the high quality reads are retained and fed into the aligner, the aligner has a large number of parameters to determine where reads originated from in the genome. These parameters are discussed at length in the STAR manual and can all be edited by the user.

Downstream, once the BED files have been generated, the MATLAB programs have a set of parameters to identify peaks and then sum the next number of values in the 5' direction. It was set that any position with less than 5 reads was cut out from being potentially listed as peak. In addition, peaks must have been 1.5 times higher than their neighbors. Lastly, after peaks were identified, summing took place 5 positions in the 5' direction.

Future direction

More libraries will be generated in the near future of Thg1 and BtTLP strains to gain statistical significance. The two samples now look very similar and in most cases make subtle difference between the RNA pools difficult to tease out. Replicates are imperative to identifying additional substrates that may only be present in low copy numbers.

RNA-seq will be performed on *Dictyostelium discoideum* (Dicty) and potentially any other TLP. In Dicty TLP4 is currently under investigation and would be knocked down with RNAi and the transcriptome compared to that of the wild type.

A genetic experiment ought to be performed with *B. thuringiensis* as well. Although previous attempts at creating a viable TLP knockout were not successful, this lends evidence that the gene is essential despite only a small number of bacteria having retained their TLP genes. Also, RNAi is not possible in bacteria as they do not possess the proper cellular machinery. Another experiment that was unsuccessfully attempted was putting the TLP gene under the influence of the Pspac promoter sequence. Thus it seems that current genetic tools are unable to probe *B. thuringiensis*. Another option is to find a different bacterial TLP to examine. Although, those bacteria that possess TLPs are not those which genetic experiments have traditionally been performed on and genetic manipulation tools are limited.

Based off the sequence of BtTLP, it is localized in the cytoplasm of yeast as there are no nuclear tags. Although, given that bacteria have no nucleus, it is of interest to see which RNAs are acted on by BtTLP in the yeast nucleus, as tRNAs are located everywhere in the cell, not just the cytoplasm. If BtTLP were engineered to have a nuclear tag, then a RNA-seq experiment could be run to identify if any RNA substrates are found in the nucleus. It is possible that these enzymes would act on precursor tRNAs in the nucleus before they could be acted on by RNase P (figure 14).

During manual examination of tRNA sequences it was noted that $tRNA^{His}$ and a small number of $tRNA^{Gln}$ appeared to be the only tRNAs being acted on differentially between control and experimental samples. BtTLP may still be adding nucleotides to different tRNAs or other RNAs though that are going undetected. If the reason for this is determined not to be from a lack of sequencing depth, then it cannot be excluded that these molecules after having nucleotide addition may be getting rapidly broken down by the cell. Such a pathway is already in known to exist with tRNAs called *Rapid tRNA decay*. Work was done showing that a *met22* deletion resulted in an accumulation of incorrectly modified tRNAs[13]. This deletion could be made in the strains used in this study to get a view of this transient transcriptional landscape.

Currently, tRNAs do not sequence well. Often times it is apparent where the reverse transcriptase comes across a modification that it cannot sequence through (figure 13b). In order to improve the ability for tRNAs to be sequenced, the isolated RNA pool could be treated with demethylases and deaminases that strip modifications off tRNAs that would otherwise halt the reverse transcriptase when making cDNA.

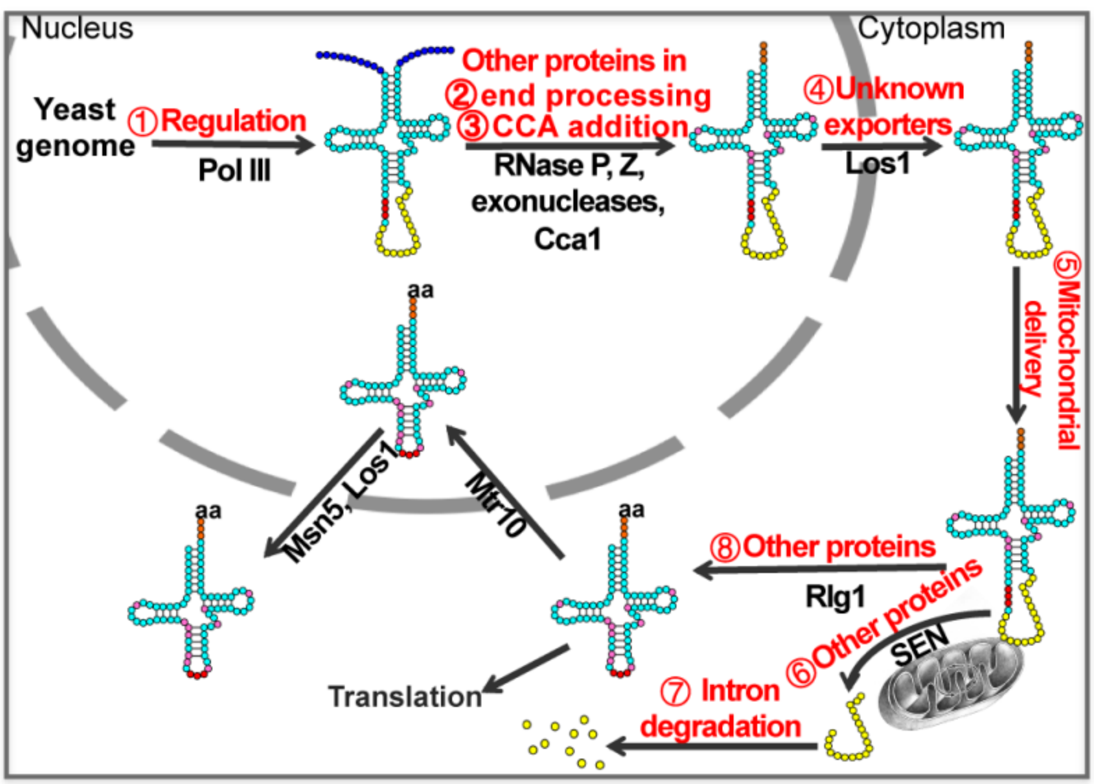


Figure 14: The tRNA synthesis pathway. tRNAs are transcribed with a leader and trailer sequences and sometimes include introns. The leader and trailer sequences are trimmed before the tRNAs leave the nucleus. If BtTLP were localized in the nucleus, it is possible that it may act on these precursor tRNAs. This figure was provided by Dr. Jingyan Wu and Dr. Anita Hopper.

Potential application

With adaptation, this pipeline could help any researcher identify post-transcriptional nucleotide addition from either the 3' or 5' ends. In a world where the central dogma seems to be less and less the case, this pipeline may prove useful in applications for studies other than reverse polymerases.

Works cited

1. Abad, M. G., Bhalchandra, S. R., & Jackman, J. E. (2010). Templated-dependent 3'-5' nucleotide addition is a shared feature of tRNAHis guanylytransferase enzymes from multiple domains of life. *PNS*, 674-679.

2. Abad, M. G., Long, Y., Willcox, A., Gott, J. M., & Jackman, J. E. (2011). A role for tRNAHis guanylyltransferase (Thg1)-like proteins from Dictyostelium discoideum in mitochondrial 5'-tRNA edition. *RNA* .

3. Alexander, D., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , 1-7.

4. Boeke, J. D., Trueheart, J., Natsoulis, G., & Fink, G. R. (1987). 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods in Enzymology* , 164-175.

5. Illumina. (2014, April 23). *Technical note: Informatics.* Retrieved March 31, 2015, from Understand Illumina Quality Scores: http://www.illumina.com/documents/products/technotes/technote_understanding_quality_scores.pdf

6. Jackman, J. E., & Phizicky, E. M. (2006). tRNAHis guanylyltransferase catalyzes a 3'-5' polymerization reaction that is distict from G-1 addition. *PNAS* , 8640-8645.

7. Jackman, J. E., Gott, J. M., & Gray, M. W. (2012). Doing it in reverse: 3'-to-5' polymerization by the Tgh1 superfamily. *RNA* , 886-899.

8. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature* , 357-359.

9. Rao, B. S. (n.d.). Diverse biological functions for 3'-5' nucleotide addition reactions: tRNA repair to tRNAHis identity. *Dissertation at The Ohio State University Graduate Program in Molecular, Cellular, and Developmental Biology* , 1-173.

10. Rao, B. S., Maris, E. L., & Jackman, J. E. (2010). tRNA 5'-end repair activies of tRNAHis guanylyltransferase (Thg1)-like proteins from Bacteria and Archaea. *Nucleic Acids Research* , 1-10.

11. Wang, Z., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews | Genetics* , 57-63.

12. Zweig, A. S., Karolchik, D., Kuhn, R. M., Haussler, D., & Kent, W. J. (2008). UCSC genome browser tutorial. *Genomics* , 75-84.

13. W., Turowski, T., & Boguta, M. (2013). An interplay between transcription, processing, and degradation determines tRNA levels in yeast. *Wiley Interdisciplinary Reviews: RNA* , 709-722.

14. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* , 2078-2079.

Acknowledgements

The graduate students Yicheng Long, Cai Chen, and Kenji Oman have been a great help. I thank SROP, the Center for RNA Biology, and the College of Engineering for funding. I also appreciate the mentorship of Dr. Ralf Bundschuh and Dr. Jane Jackman, who have been patient and promoted a positive learning environment.